Prediction on Weather Insurance Purchasing Decision based on Lasso and Random Forest

Tianhao Cao

June 5, 2022

1 Introduction

How to make people purchase the product and the service is an eternal topic that every trade-involved industry needs to think about. On one side, the marketing field developed millions of methodologies to solve the issue of how to sell more products and services to customers, including advertisement, build-up users' community, raising the brand's popularity, etc. On the other side, Economists started to link the purchasing decision with customers' traits and characteristics, like social network influences, coefficient of relative risk aversion, age, education level, etc. Although both sides merged well and worked perfectly, it was then followed by a question: How to quantify the effects of all these factors that may influence the purchasing decision, and more importantly, how to predict new customers' purchasing probabilities based on the above estimations?

This essay is constructed in two parts; both will focus on predicting insurance buy deci-

sions for Chinese farmers. Using the data [CDJS18], I will analyze the significant variables that lead to huge influences on the predicted variable: insurance-buy-decision. For the first part, the main estimation method will be Lasso Cross-validation. I will test both the In-Sample deviance and the Out-of-Sample deviance to quantify the model's performance. For the second part, the main method is the Random Forest, where I will also provide In-Sample deviance as well as Out-of-Sample deviance to measure the performances. By comparing these results, I am able to determine the best model to predict the outcome $PurhcaseProbability$.

The result turns out that I failed to construct models that accurately predict the purchasing probability of farmers, where all the models have low R^2. However, I found some essential factors correlated with the purchasing probability while constructing the models.

2 Literature Review

Topics about identifying significant variables that influence the probability of purchase decisions have been widely discussed in multiple areas. [AD10] discussed the social network effects on alcohol consumption among teenagers, and it concludes that "a 10% increase in the proportion of classmates who drink will increase the likelihood of drinking participation and frequency by approximately four percentage points." their OLS estimation indicates that social network has the greatest influence on individuals decision to consume alcohol. Same with the article [CFS 8] illustrates that "There is a sizable share of the market (38 percent) that can be positively impacted through social network marketing". Contrarily, the original article of the data [CDJS15] indicates that social network has "no significant

effect of friends' decisions on individuals' choices." (Jing Cai et al. 2015, Page 82). This is because the treatment D has selection bias, which in the article, they conclude, "The main mechanism through which social networks affect decision making is social learning about insurance benefits, as opposed to the influence of friends' purchase decisions which are not transmitted in social networks." (Jing Cai et al. 2015, Page 82). My result turns out both *Understanding* and Network-related variables are important factors that influence a farmer's buying decision, but *Understanding* weighs more than social-network. The intuition behind this is that weather insurance is a different product compared to other goods and services, which regular goods and services will lead to returns, such as alcohol, food, and games. However, in terms of insurance, it sometimes leads to no returns.

3 Data

3.1 Data Introduction

The data is found on the Harvard Database; the article's author posted four meta-data and the corresponding .do file, which allows me to recurrent the merged dataset. The data is a low-dimensional dataset containing 4902 observations and 59 variables, which involved 5300 households chosen from 185 rural villages that were randomly selected in Jiangxi, China. Each observation represents a household that participated in the experiment. Since the observations are randomly selected, I believe the data has good representativeness in predicting rural farmers' insurance-buy-decision within Jiangxi in 2010.

3.2 Data Cleaning

However, the first-hand data contains large numbers of NA values under several variables. Hence, I dropped several variables based on the criteria: the number of NA values is more than 1000 in an effort to maximize the number of observations while cleaning out all the NA values. Including $DisasterLoss$, a continuous variable that measures the loss in yield due to disasters last year; $Reveal$, a binary variable that measures the whether the observation revealed purchase decisions to Friends; $NetworkRatePretakeup$, a continuous variable measures the proportion of the observations who attend the first round insurance session and buy the insurance; $KnowledgeNetwork$, a continuous variable measures the mean understanding of those people who were not assigned to second round sessions; Other neural network related variables that original author created.

3.3 Data Limitations

The data itself and the cleaning process create bias. Firstly, my estimation will be biased due to the data cleaning process, some important variables are dropped to maintain enough observations, and such a process will lead to biased results. Secondly, the analysis of the dataset may be implausible under today's scenario because of some omitted variables that I can not observe. In the first place, technologies are more advanced to detect natural disasters and protect the rice field from being damaged by natural disasters, and the risk-averse of the observations have changed. Meanwhile, Internet and mobile phone technologies are more advanced, and farmers are more informed through different channels; in other words, their understanding of insurance is more likely to vary from the sample of the data, which is the

second difference. Also, the education level may change a lot compared to 2010. The age variable in the table1 shows that the median age of the population had already been 50 years old, which they may have retired right now. Meanwhile, new farmers' education levels are expected to be higher than the sample in the dataset, which leads to the third difference compare with current sample.

3.4 Overview data

Table 1: Summary Statistics based on Region

region		1				2				3		
Variable	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median	Max	Min
day	13.6	14	25	1	28.1	25	51	18	34.9	34	40	29
risk_averse	0.2	0	1	0	0.2	0	1	0	0.1	0	1	0
insurance_repay	0.3	0	1	0	0.5	1	1	0	0.5	0	1	0
age	52.4	52	88	20	48.4	46	83	19	50.3	49	82	20
educ	1.2	1	4	0	1.3	1	4	0	1.2	1	4	0
understanding	0.4	0.4	1	0	0.4	0.4	1	0	0.5	0.4	1	0

$$\text{Understanding} = \beta_0 + \beta_1 * \text{day} + \mu \tag{1}$$

$$\begin{aligned}\text{Understanding} = {} & \alpha_0 + \alpha_1 * \text{day} + \alpha_3 * \text{region2} + \alpha_4 * \text{region3} + \alpha_5 * \text{day} * \text{region2} \\ & + \alpha_6 * \text{day} * \text{region3} + \epsilon\end{aligned} \tag{2}$$

Table 1 indicates the summary statistics of different regions, my interests of estimation are $TakeupSurvey$ as the outcome Y; $NetworkObs$, $Understanding$, $educ$, $risk_averse$, and all other remaining variables as regressors. This graph illustrates the magnitudes of different variables. As we can see, all three regions have multiple identical variables, such as average

age, education, disaster probabilities, etc. However, the rice income and rice area in 2010 of region 2 are higher than in the other two regions. Also, I notice that the average days of sessions are steadily increasing through region 1 to region 3 without an increase in average Understanding.

Table 2: Effects of Experiment days on Understanding

	Dependent variable:		
	understanding		
	(1)	(2)	(3)
day	0.001** (0.0004)	0.001 (0.001)	0.001 (0.001)
region2		0.054 (0.034)	0.033 (0.068)
region3		−0.134 (0.122)	−0.205 (0.137)
risk_averse			0.065*** (0.019)
insurance_repay			−0.058*** (0.013)
age			−0.001* (0.001)
educ			0.055*** (0.008)
day*region2		−0.002* (0.001)	−0.001 (0.001)
day*region3		0.004 (0.004)	0.007* (0.003)
region2*risk_averse			−0.008 (0.036)
region3*risk_averse			−0.038 (0.042)
region2*insurance_repay			0.051** (0.022)
region3*insurance_repay			0.070*** (0.025)
region2*age			−0.001 (0.001)
region3*age			−0.0005 (0.001)
region2*educ			0.022 (0.013)
region3*educ			0.003 (0.015)
Constant	0.421*** (0.010)	0.416*** (0.013)	0.425*** (0.037)
Observations	4,166	4,166	4,166
Log Likelihood	−825.061	−820.640	−705.319
Akaike Inf. Crit.	1,654.121	1,653.281	1,446.638
Note:			*p<0.1; **p<0.05; ***p<0.01

Hence, I estimated the effects of experiment days on Understanding with the above regressions. Formula (1) is the pure correlation between experiment days and outcome: Understanding, whereas formula (2) indicates the correlations of Understanding with the interactions between experiment days and regions. Also, I include several other variables in third regressions just to provide enough information. As we see that after including

interactions and other variables, experiment days become statistically insignificant, I can conclude that experiment days are not statistically influencing the understanding of Chinese farmers about the weather insurance.

4 Method

4.1 Lasso Cross-validation

Lasso is a regularization method to minimize the deviance of ordinary least squared by adding penalties to the OLS estimation, and the penalty is controlled by λ. The way of choosing λ will then be controlled by the cross-validation method. In this empirical analysis, I cross-validate the model with $n = 30$, then test both In-Sample deviance and Out-of-Sample deviance for the best model as well as classify the prediction based on these models. The final result will be shown in ROC Curves. To use the Lasso model because, although low-dimensional data is applicable for OLS, I can not determine the best complexity for the model. Due to the variance bias trade-off, I need to test as many as OLS regressions to find out the optimal complexity, but in Lasso, I am able to use cross-validation to determine the optimal λ to use; this highly decreases my workload as well as highly improves both the performance and the credibility of the model.

4.2 Random Forest

The random Forest method is a way to better evaluate In-Sample deviance and Out-of-Sample deviance through averaging over-fit predictions in different regression trees. The

reason why I also include Random forest as one of the methods is that Lasso has a disadvantage, which is I can't find the best regression at first glance; this indicates that the OLS model or Lasso model is still limited by the fact that I don't know the true correlations between regressors and outcomes. Random Forests is able to find out the correlations between regressors and outcomes without manually inputting them. Although it requires larger numbers of observation to estimate, I am able to fake the dataset using bootstrapping, which ideally solve the issue that the dataset lacks observations. Based on the above, I am able to predict the outcome from the Random Forest method to compare it with the Lasso CV method to see which one works better.

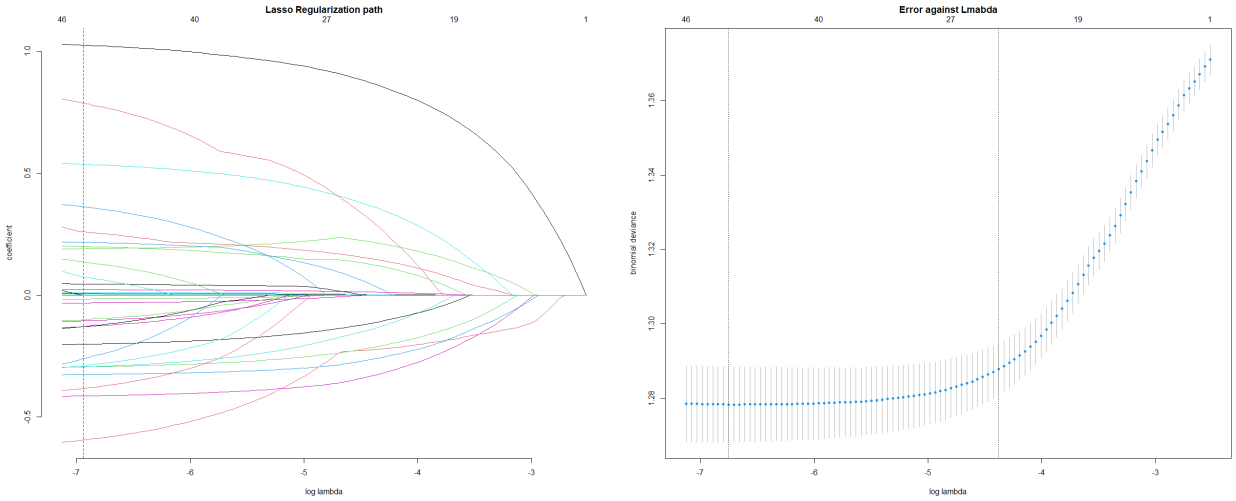
4.3 Classification Criteria

Both the Lasso Cross validation and the Random Forest methodologies will be classified based on natural cut off value: $p = 0.5$. Additionally, I will provide different cut off values $p = 0.02, 0.1, 0.5, 0.8, 0.9$ within ROC curves.

4.4 Representativeness

Both Lasso Cross-validation and Random Forest require representative samples to estimate the true correlations between dependent variables and regressors. The outcomes that they predict depend on these correlations estimated. Hence, although the data itself has good representativeness to some extent and models are estimated based on the data, models can not use them to predict current farmer purchasing decisions due to the reasons that I mention in section 3.3 Data Limitations. However, these models can not only be used to predict, for

Figure 1: Lasso Regularization path and Error Against Lambda



example, 2011 Jiangxi farmers' weather insurance purchasing decisions, but we can also use these models to predict weather insurance purchasing decisions for those farmers who live in Laos[1], Thailand[2], and those countries have a similar background[3] with Jiangxi in 2010.

5 Results

5.1 Lasso Cross-validation

Figure 1 indicates the Lasso Regularization Path and Error against Lambda of the Lasso model, I include all the regressors except *Village*, *Address*, and *InsuranceBuy*, where the first two regressors indicate the village and address index for each observation and the third regressor is the prediction on the purchasing decisions made by the original author of the dataset. There are 4166 observations and 55 variables included in this estimation, and the optimal lambda zeroed out 14 of the regressors, which the Lasso model with 41

[1]The main food crop in Laos is rice, which accounts for 85 percent of the country's crop area.
[2]Thailand is the world's largest rice exporter
[3]Technology, GDP, Main food crop

regressors contains optimal variance and bias. Among these 46 regressors, $Understanding$, $NetworkTwoside$, $Region1$, $RiskAverse$, $default0$, $Region2$, and $NetworkObs3$ are the seven regressors that have largest magnitudes.

5.1.1 In-Sample Estimation

Figure 2 indicates the distributions of predicted probabilities and the ROC curve for In-Sample predictions. The distribution of the probabilities most lies in the middle of the distribution graph, which indicates that the Lasso model is uncertain about the predictions. Furthermore, the ROC curves illustrate the trade-off between sensitivity and specificity. The sensitivity is equal to 0.43, and the specificity is equal to 0.78 when $p = 0.5$, the rest of the cut-off points either have incredibly high specificity or extremely high sensitivity. Additionally, I measure the correctness of the predictions based on the collected purchasing decisions. It turns out that 64.07% of my predictions match the actual purchase decisions. Although the correctness is delightful, the R^2 is only 0.0788, which indicates the model is not precise and accurate.

5.1.2 Out-of-Sample Estimation

In Out-of-Sample Lasso estimation, I randomly split the data into two equal parts: the training set and the test set. To re-predict the test set's outcome based on the training set's estimation, the outcome of the model is $R^2 = 0.0713$. Figure 3 indicates the distributions and ROC curves for Out-of-Sample predictions, and the correctness of the predictions 63.37% match with the actual purchase decision. As we can see that the results are quite similar, both In-Sample and Out-of-Sample estimations do not result in a good fit to the dataset.

Figure 2: In-Sample distribution and ROC curve

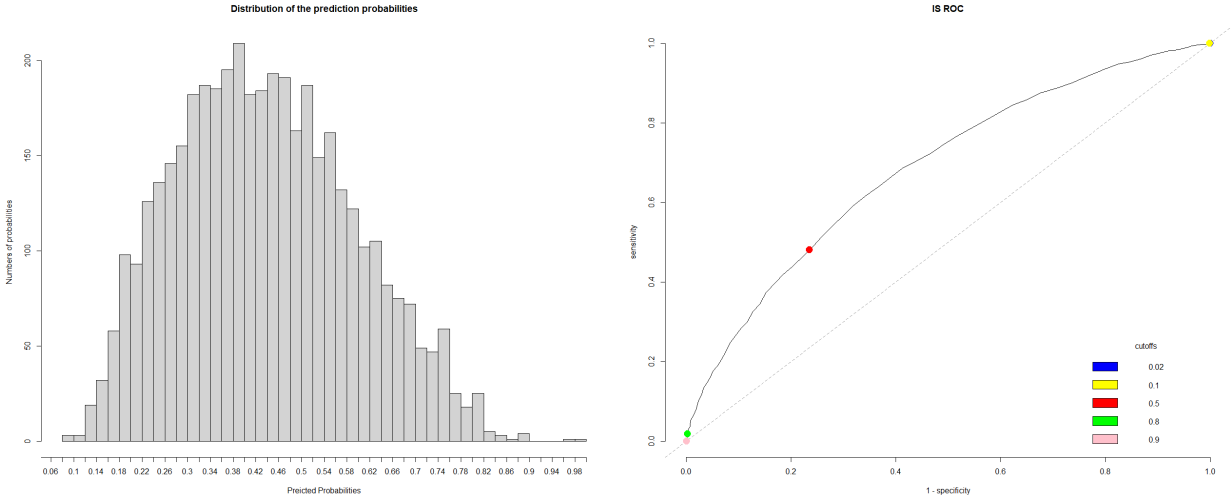


Figure 3: Out-of-Sample distribution and ROC curve

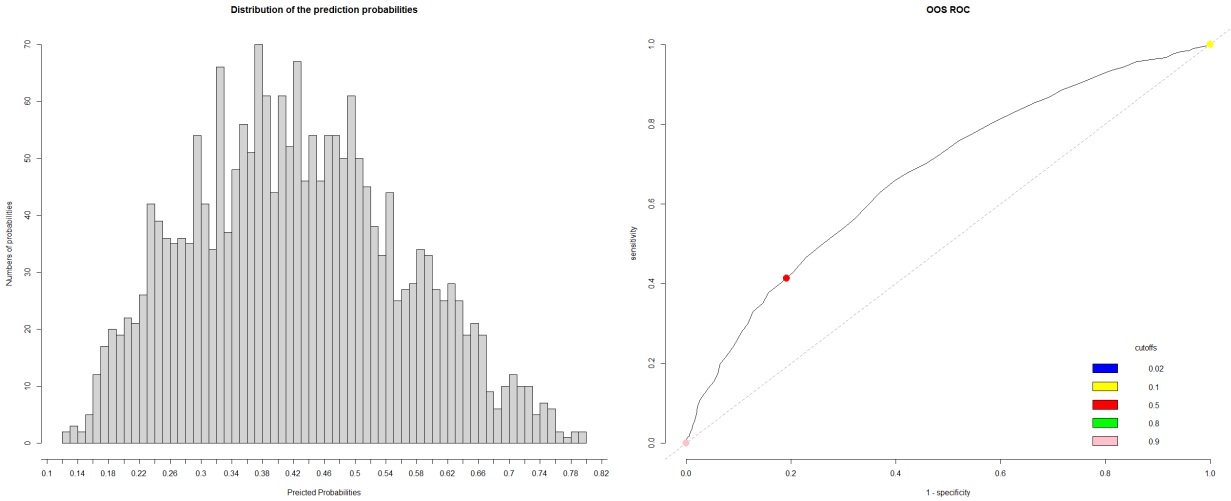
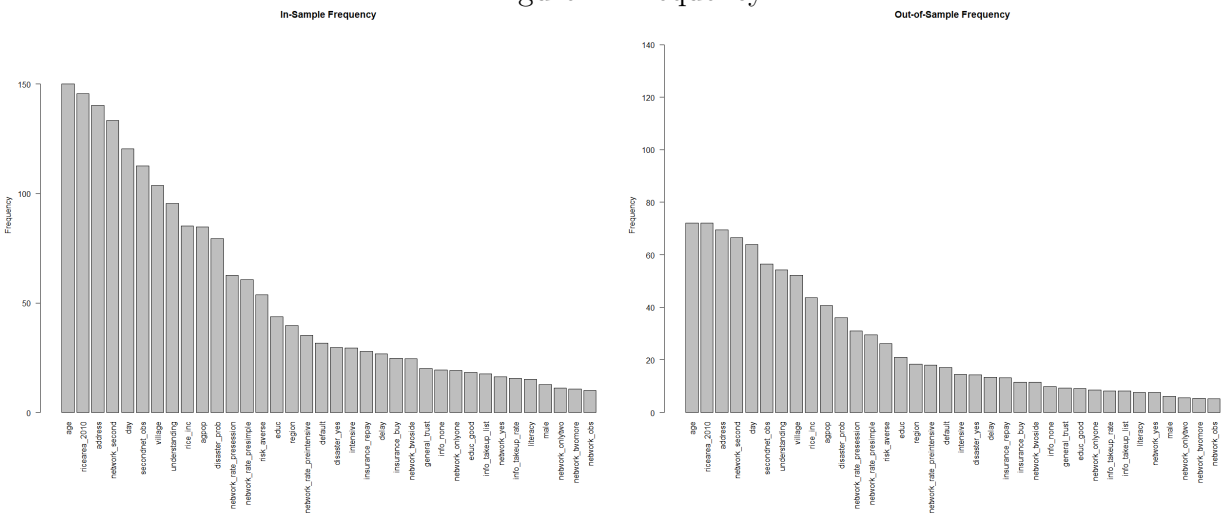


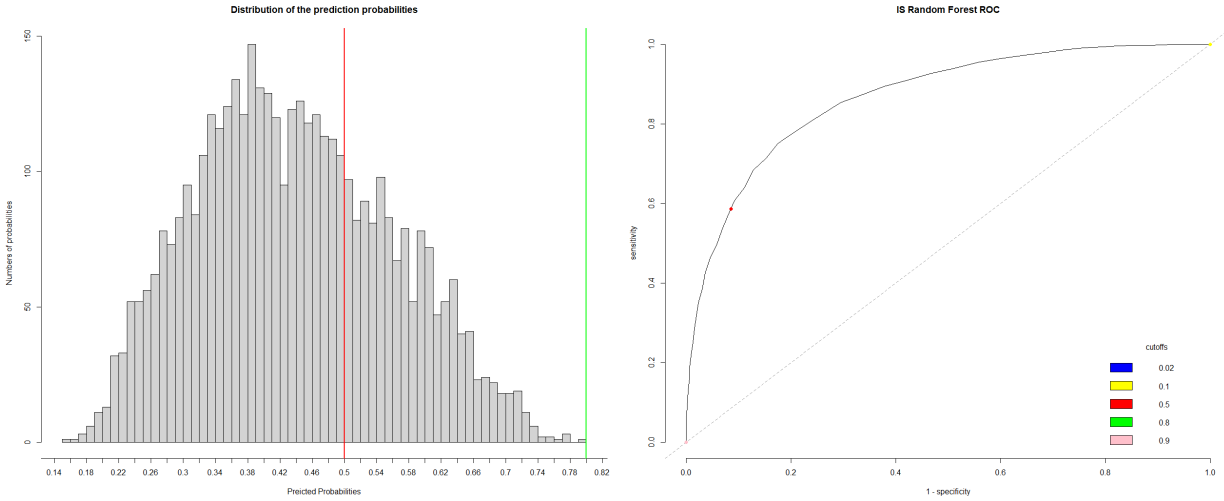
Figure 4: Frequency



5.2 Random Forest

The random forests of the estimations are constructed with 300 trees, and each tree has a minimum node size of 100. Due to the characteristic of the Random Forest algorithm, I can not get the exact correlations between regressors and independent variables, but I am able to extract the importance of different variables. Figure 4 indicates the In-Sample and Out-of-Sample Random forest frequency for the estimations, where the frequency of the variable indicates the importance of the variable. The Random Forest algorithm gives out a similar result to Lasso; the different part is that *age* has a considerable influence on outcome in Random Forest, and observations' risk traits are not as important as they are in Lasso. A similar part is that *Understanding* is the most important variable to estimate the outcome, and network-related variables, are significant in influencing the outcome. Surprisingly, *RiskAverse* is not an important factor that influences the purchasing decision in the Random Forest algorithm. From the common economics concept, farmers should tend to buy insurance to avoid the loss brought by disaster if they have a higher risk-averse rate.

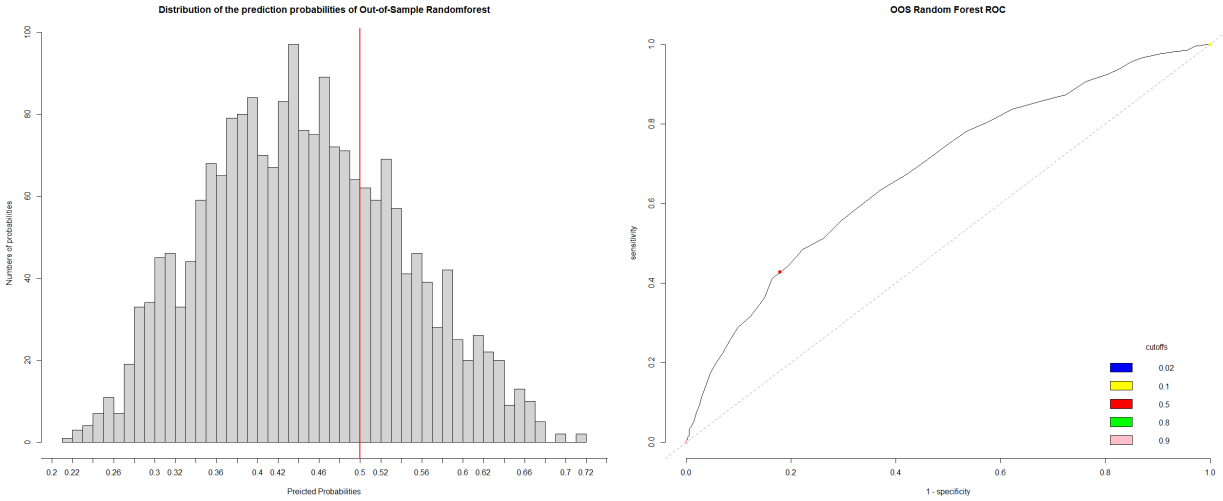
Figure 5: In-Sample distribution and ROC



5.2.1 In-Sample Random Forest

Figure 5 indicates the In-Sample distribution of probabilities and the In-Sample ROC curves of the random forest algorithm. On the one hand, the random forest algorithm has a better shape ROC curve than the Lasso In-Sample ROC curve, where the specificity = 0.9 and sensitivity = 0.1 when $p = 0.5$. On the other hand, the distribution of probabilities is similar to the Lasso model, which mostly lies in the middle of the distribution graph, indicating that the algorithm is also uncertain about the majority of the predictions. I tested the correctness of the prediction, and it shows that 77.12% matches with the actual purchasing decision. However, the In-Sample Random Forest model has the $R^2 = 0.1847$, which I believe is that the model is overfitted to the data based on the performances of other models. The data itself lack critical regressors that are highly correlated with the outcome, and such a high R^2 is impossible.

Figure 6: Out-of-Sample distribution and ROC



5.2.2 Out-of-Sample Random Forest

The Out-of-Sample Random forest model has the $R^2 = 0.0627$, which is also a deficient precision model, just like Lasso. The ROC Curve also proved that the Out-of-Sample predictions of the Random Forest result in a bad fit to the model due to these lowly correlated regressors. The Random Forest ROC curve and Lasso ROC Curve are pretty identical in that they both have bad fits to the dataset, where the specificity and the sensitivity are highly trade-offs.

6 Conclusion

To sum up, the essay aims to accurately predict farmers' weather insurance purchasing probabilities using Lasso and Random Forest. However, due to the limitation of the dataset, I fail to estimate the true correlations, which the R^2 of all the models turns to be less than 0.1. My understanding of such a phenomenon is that the data contains a few highly correlated variables that influence the outcome, such as the price of the insurance, the income of the

household, is the rice income the most important income for the family, so on. Nevertheless, it does not mean that the data is not representative, I am able to know the several significant variables that influence the purchasing probabilities through the retained variables from the regularization path of the Lasso as well as the variable frequencies of Random Forest. The above models indicate that, same as the original article, $Understanding$ is a highly influential variable during the estimations in Lasso Cross-Validation and Random Forest. Meanwhile, network-related variables also took important roles during estimations. Additionally, several other variables have different weights in different algorithms, such as age, $RiceArea$, Day, $RiskAverse$, $Region$, and $default$, and the study of the essay may be helpful to the future study of such an area.

References

[AD10] Mir M. Ali and Debra S. Dwyer. Social network effects in alcohol consumption among adolescents. *Addictive Behaviors*, 35(4):337–342, 2010.

[CDJS15] Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

[CDJS18] Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Replication Data for: Social Networks and the Decision to Insure, 2018.

[CFS 8] Colin Campbell, Carla Ferraro, and Sean Sands. Segmenting consumer reactions to social network marketing. *European Journal of Marketing*, 48(3/4), 2014-4-8.